DOCUMENT RESUME

ED 306 303 TM 013 194

AUTHOR Robertson, Gary; And Others
TITLE The Refinement of a Single Scale for the Measurement of Writing Ability.
PUB DATE Mar 89
NOTE 13p.; Paper presented at the Annual Meetings of the American Educational Research Association (San Francisco, CA, March 27-31, 1989) and the National Council on Measurement in Education (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary School Students; Elementary Secondary Education; Interrater Reliability; *Latent Trait Theory; *Pictorial Stimuli; *Rating Scales; Regression (Statistics); *Scaling; Secondary School Students; *Test Construction; Test Reliability; Timed Tests; *Writing Evaluation
IDENTIFIERS Graded Response Model; *Peabody Individual Achievement Test (Revised); Writing Samples

ABSTRACT

A study was conducted to refine the writing scale incorporated into the Peabody Individual Achievement Test-Revised. The test uses a single scale for judging writing samples from students in grades 2 through 12. It was questioned whether a single, relatively brief rating scale could have the sensitivity required to discriminate among the range of written expression presented by students across this broad grade range. The written expression subtest provides the students with 20 minutes to write a story in response to one of two picture prompts. Each of 24 discrete items is rated on either a dichotomous or trichotomous scale, depending on the characteristic being rated. Each item on the scale yields a point score. For this study, 1,087 writing samples were collected from students in grades 2 through 12 in the spring of 1986. About half of the students took Prompt A, and the other half took Prompt B. A subset of 168 students was selected to retake the alternate prompt for later use in estimating alternate-forms reliability. Item weighting methods applied to the data include regression, graded-response, and teachers' rating methods. Ten secondary school English teachers rated the 24 written expression items in terms of their importance to the evaluation of a student's written composition. Application of these three methods failed to yield improvements either in mean raw score growth from grade-to-grade or in internal consistency reliability. Five data tables are presented. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306303

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GARY J. ROBERTSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE REFINEMENT OF A SINGLE SCALE FOR THE MEASUREMENT OF WRITING ABILITY

Gary Robertson, Elizabeth Rengel, Jing-Jen Wang

American Guidance Service

Paper presented in in G. H. Roid (Chair), Issues in Measuring Growth and Change in Direct Writing Assessment. Symposium conducted at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education.

San Francisco, CA
March 27-31, 1989

TM013194

BEST COPY AVAILABLE

Refinement of a Single Scale for the Measurement of Writing Ability

Gary Robertson, Elizabeth Rengel, and Jing-Jen Wang

American Guidance Service

The procedure used to develop a single scale for measuring performance on the Written Expression subtest of the *Peabody Individual Achievement Test--Revised* was described in an earlier paper (Robertson, Wang, & Rengel, 1988). Picture prompts were used in grades 2-12 as the stimuli for eliciting samples of direct narrative writing skill. A single 24-item rating scale was developed to assess important aspects of written expression; each item was rated on a scale ranging from 0 to 2 points. Scoring criteria were provided for assigning each of the rating scale values used with each of the rating scale items.

Total raw scores on the 24-item rating scale were obtained for all students in the standardization sample, grades 2-12. Although the psychometric characteristics of the single rating scale were satisfactory for a wide-range screening measure such as the *Peabody Individual Achievement Test--Revised*, certain problems were noted which precluded the development of within-grade standard scores. Mean raw scores showed only a slight increase beyond grade 6; coefficient alpha reliabilities, while acceptable, were less than ideal. Inter-rater reliability coefficients constituted more of a problem, with median values around .6 and .7 for the two writing prompts. There was, however, considerable variability in inter-rater reliability across grades due partly to small sample size.

Purpose of the Present Study

There was a recognition from the outset that the use of a single scale for judging writing samples in grades 2-12 presented certain problems. One fundamental question centered around the extent to which a single, relatively brief rating scale would have the sensitivity required to discriminate among the range of written expression evidenced in writing samples of students in grades 2-12. A related concern was whether meaningful

interpretations could be generated from such a brief scale. These questions had not been answered by previous research because the use of a single rating scale with the PIAT-R was the first instance of such an application on a large-scale basis.

The present study was undertaken in an attempt to refine further the scale by use of three item weighting schemes designed to maximize the contribution of each item to total score according to certain criteria. Each item weighting procedure is described briefly in the following section. The PIAT-R Written Expression data were analyzed to determine the effect of these item weighting procedures on raw score parameters and reliability estimates across grades 2-12.

Method

Instrument

The PIAT-R Written Expression subtest was designed to assess the writing skill of students in grades 2-12. The student is given 20 minutes to write a story in response to one of two picture prompts (Prompt A or B). The scoring system is based on a procedure suggested by Cohen (1973): a set of 24 discrete items is rated on either a dichotomous or a trichotomous scale, depending on the characteristic being rated. The items are similar to the following example:

	(Circle the appropriate value)
Run-on sentences are avoided.	2 = Almost always
	1 = Sometimes
	0 = Rarely

Each item on the rating scale yields a point score. The scores for all items in the scale may be summed to obtain a total score. Scaled scores comparing performance with the total standardization sample and within-grade stanines are available for interpretive purposes. The development and technical characteristics of the Written Expression subtest are fully described in the PIAT-R manual (Dunn & Markwardt, 1989).

Subjects

A total of 1,087 writing samples was collected from students in grades 2 through 12 during the national PIAT-R standardization

program in the spring of 1986. Approximately half of the students took Prompt A, and the other half took Prompt B. A subset of 168 students was selected to retake the alternate prompt for later use in estimating alternate-forms reliability. In addition, each writing sample was scored by two trained raters in order to provide data for estimating inter-rater reliabilities.

Data analyses

The following three differential item weighting methods were applied to the standardization data:

Regression method. Item score was regressed on grade, and the resulting correlation was rescaled to provide the item weight. This method gives more weight to items that correlate higher with grade level, and less weight to those with lower correlations.

Graded response model. The graded response model (Samejima, 1969) with two parameters (difficulty and discrimination) was the second differential item weighting procedure used. The data from Prompts A and B were calibrated separately by the computer program MULTILOG (Thissen, 1984). Estimates of the item discrimination parameters from the calibration results were rescaled and used as item weights.

Teachers' rating method. Ten secondary English teachers rated the 24 Written Expression items in terms of their importance to the evaluation of a student's written composition. The teachers used the following five-point rating scale:

5 = most important in good writing

4 = highly desirable but less important than #5

3 = moderately important

2 = less important than #3 but not least important

1 = least important

The teachers rated the importance of the Written Expression items separately for each of three grade groups: 2-4, 5-8, and 9-12. The mean rating for each item was rescaled and used as the item weight.

Inter-rater reliability of the teachers' ratings was assessed by Ebel's (1951) method. The reliabilities were .86 for grades 2-4, .85 for grades 5-8, and .83 for grades 9-12.

The three differential weighting methods were applied to the standardization data and means and standard deviations were computed. In addition, reliability coefficients (alpha, inter-rater) were obtained for each weighting method. These data were used to compare the weighting methods with each other and with the original unweighted data to evaluate the effects of differential item weighting.

Results

Differential item weights for the three weighting procedures are shown in Table 1. Standard deviations were also computed to show the dispersion of item weights within each method. Table 1 shows that the regression method and the graded response model produced a greater dispersion of weights than did the teachers' rating method which is based on a narrower span of grades. All three methods have dispersion values greater than zero--the dispersion value that can be expected from the unweighted method.

Insert Table 1 about here

Raw score means and standard deviations by prompt and grade for the unweighted and the three weighted procedures are reported in Table 2. In general, the mean raw scores (both unweighted and weighted) show the expected negatively accelerating growth

---

Insert Table 2 about here

---

progression across the grade span, although the growth is far from the smooth progression characteristic of the other PIAT-R subtests and most other academic achievement tests. In addition, there is a tendency for the Written Expression scores to reach an asymptote earlier than the other PIAT-R subtests. This finding is consistent with findings reported by Wiser (1988) and Myklebust (1965) on other writing assessment instruments.

To investigate the effects of differential item weighting on mean raw score progressions across the grade span, the average increment of consecutive grade mean raw scores is also reported in Table 2. The average mean progression across grades is 1.48 for the regression method and 1.39 for the graded response model. Both values are greater than that of the unweighted method (1.19). Surprisingly, the average mean progression for the teachers' rating method (1.08) is smaller than that of the unweighted method (1.19). In general, although there are some improvements in mean raw score progression across the grade span for the regression and graded response methods, the lack of growth in the higher grades (8-12) remains unchanged.

Also of interest is the effect of differential item weighting on test reliability. Three reliability estimates were obtained and reported in Tables 3 through 5. Results in Table 3 clearly

---

Insert Tables 3,4, and 5 about here

---

indicate that applying any differential item weighting method to the test has no effect on coefficient alpha. The item weighting procedures produce a slight improvement in inter-rater reliabilities (Tables 4 and 5). The graded response method shows the largest increase in inter-rater reliability calculated by either the conventional method or by Ebel's (1951) method. The lack of improvement in reliability is not totally surprising. Similar results have been reported in previous researches on differential item weighting (Wang & Stanley, 1970).

Discussion

This attempted refinement of the PIAT-R Written Expression raw score scale through the application of differential item weighting was undertaken because it was felt that the use of a single rating scale for all grades, grades 2-12, with equal weight given to all items might restrict grade-to-grade increments in raw score as well as internal consistency and inter-rater reliability estimates. The lack of improvement in grade-to-grade raw score increments and internal consistency after weighting presents three interesting possible explanations for consideration.

First, the lack of improvement after weighting may be an artifact of the weighting procedure itself; the method may well not have the sensitivity required to compensate for the fact that grade-to-grade mean raw score increments are small. Related to this is the further possibility that the methodology used to elicit and evaluate the direct writing samples is not sensitive enough to register growth in written expression that may actually occur developmentally. A third possible explanation relates to the nature of the growth in written expression skills across the grade levels: there may actually be relatively little growth after grades 6 or 7, at least in narrative writing. Wiser (1988) reported a result similar to those reported here with the *Stanford Achievement Test Writing Assessment*. It may well be that the growth in written expression skills across the elementary and secondary grades follows a pattern different from that shown by other content disciplines such as reading and mathematics, for example; moreover, there may be less variability in written expression skills than is true of achievement in other academic areas.

The generally observed lack of effect of weighting on mean raw score increments and reliability would argue against any attempts to develop a hand-weighting procedure for routine use with the test. The widespread availability of microcomputers frees the test user from much of the clerical burden imposed by a hand weighting procedure. The slight increase in the graded response model inter-rater reliabilities may justify the application of a weighting procedure when microcomputer scoring is available. Additional investigation is needed, however, before this step is undertaken by users.

Conclusion

The application of three different item weighting schemes to the 24 rating scale items used to evaluate PIAT-R direct writing samples for the standardization examinees failed to yield improvements either in mean raw score growth from grade-to-grade or in internal consistency reliability. The graded response two-parameter IRT model did produce slight increases in inter-rater reliability. Further evidence is needed before item weighting is applied to augment the results obtained from direct writing assessments which use this particular rating scale methodology.

References

Cohen, A. M. (1973). Assessing college students' ability to write compositions. Research in the Teaching of English, 7, 356-371.

Dunn, L. M., & Markwardt, F. M. (1970). Peabody Individual Achievement Test Manual. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Markwardt, F. M. (1989). Peabody Individual Achievement Test--Revised. Circle Pines, MN: American Guidance Service.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Myklebust, H. (1965). Development and disorders of written language: Volume One, Picture Story Language Test. New York: Grune & Stratton.

Robertson, G. J., Wang, J. J., & Rengel, E. (1988). The development of a single writing scale for assessing writing skills in grades 2-12. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, No. 17.

Thissen, D. (1984). MULTILOG [Computer program]. Mooresville, IN: Scientific Software.

Wang, M. & Stanley, J. (1970). Differential weighting: a review of methods and empirical studies. Review of Educational Research, 4, 663-705.

Wiser, B. (1988). Building a developmental scale for a nationally normed writing test. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Table 1
Differential Item Weights for Three Weighting Procedures

	Prompt A		Prompt B		Teachers' Ratings[a]		
Item	Regression	Graded Response	Regression	Graded Response	Gr. 2-4	Gr. 5-8	Gr. 9-12
1	.587	.730	.213	.708	1.234	.990	.944
2	.820	.996	1.024	1.169	1.332	1.150	1.156
3	.953	1.106	.828	.981	.974	.990	1.015
4	.766	.732	.612	.676	.942	.990	1.038
5	.941	1.087	1.250	1.096	.877	.990	1.062
6	.929	.404	.836	.641	.812	.910	.991
7	1.408	1.059	1.530	1.276	.974	.990	1.038
8	1.106	.864	1.256	1.132	1.332	1.258	1.109
9	.813	.824	1.036	1.013	1.234	1.231	1.086
10	.879	1.088	.400	.756	1.137	1.204	1.109
11	.992	.613	.901	.565	1.007	1.017	.920
12	.888	.563	.977	.478	.747	.776	.897
13	1.620	1.345	1.821	1.368	1.234	1.124	1.109
14	1.080	1.419	.984	1.733	1.559	1.284	1.180
15	.963	.787	.922	.689	1.137	1.097	1.062
16	.956	.388	.802	.433	.682	.990	1.062
17	.617	.707	.267	.582	.909	.963	.968
18	1.240	1.601	1.426	1.472	1.202	1.070	1.109
19	1.402	2.211	1.501	1.589	1.234	1.124	1.086
20	.844	.612	1.064	.745	.942	.910	.850
21	1.221	1.020	1.385	1.353	.877	.883	.897
22	1.237	1.354	1.121	1.097	.585	.696	.732
23	.987	1.489	1.276	1.447	.584	.669	.732
24	.750	1.000[b]	.568	1.000[b]	.455	.696	.850
SD	.247	.411	.394	.366	.268	.169	.121

Note: The weights are rescaled such that the sum of weights for a method is equal to 24.
[a] Teachers rated the 24 scoring criteria; they did not rate separately by prompt.
[b] A weight of 1.000 was assigned because the calibration program could not obtain convergent estimates of the item parameter.

Table 2
Means and Standard Deviations for Unweighted and Three Differential Item Weighting Procedures, by Prompt and Grade

	Prompt A								Prompt B							
	Unweighted		Regression		Graded Response		Teachers' Ratings		Unweighted		Regression		Graded Response		Teachers' Ratings	
Grade	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
2	26.7	9.5	25.2	9.9	24.1	10.0	28.6	9.9	27.2	8.0	24.4	9.4	25.0	9.1	29.0	8.4
3	31.0	6.8	30.0	7.4	29.1	7.9	33.1	7.1	28.9	9.8	26.9	11.1	27.3	10.7	30.9	10.3
4	32.0	8.1	31.1	8.7	29.9	8.9	34.1	8.1	30.5	8.9	29.0	10.1	29.3	9.8	32.5	9.0
5	35.3	6.9	34.9	7.6	34.3	8.4	36.6	6.7	37.2	5.9	37.0	7.0	37.0	6.7	38.5	5.6
6	35.3	7.2	34.5	8.3	33.7	8.9	36.7	7.2	36.2	6.2	35.6	7.4	35.6	7.1	37.4	6.1
7	36.8	5.2	36.3	5.8	35.2	6.5	38.2	4.9	35.8	7.7	35.0	9.2	35.0	8.9	37.1	7.5
8	38.5	6.1	38.4	6.7	37.9	7.6	39.6	5.9	37.7	6.9	37.4	8.3	37.1	8.0	38.9	6.7
9	39.9	5.5	39.7	6.0	38.9	6.9	40.6	5.2	38.3	6.6	38.1	7.9	37.9	7.5	39.0	6.4
10	37.2	7.1	36.8	7.7	35.7	8.4	37.9	7.0	38.4	6.2	38.4	7.2	38.1	7.1	39.1	6.0
11	40.2	5.8	40.0	6.5	39.4	7.1	40.8	5.7	38.2	6.6	38.0	8.0	37.8	7.7	38.9	6.4
12	39.2	6.0	39.2	6.4	38.3	7.1	39.9	5.8	39.1	8.3	39.2	9.2	38.9	9.0	39.8	8.3
Mean Increment	1.25		2.67		1.42		1.13		1.19		1.48		1.39		1.08	

Table 3
Coefficient Alpha for Unweighted and Three Differential Item Weighting Procedures, by Prompt and Grade

	Prompt A				Prompt B			
Grade	Unweighted	Regression	Graded Response	Teachers' Ratings	Unweighted	Regression	Graded Response	Teachers' Ratings
2	.89	.89	.88	.90	.84	.84	.85	.85
3	.77	.77	.76	.78	.90	.88	.89	.90
4	.85	.85	.83	.86	.87	.86	.86	.87
5	.81	.81	.81	.81	.76	.76	.77	.75
6	.82	.82	.82	.82	.76	.77	.77	.77
7	.69	.70	.69	.70	.86	.85	.85	.86
8	.78	.78	.79	.79	.82	.82	.83	.82
9	.77	.77	.76	.77	.82	.82	.82	.82
10	.81	.81	.81	.81	.80	.79	.80	.80
11	.78	.79	.79	.80	.81	.82	.83	.82
12	.79	.78	.77	.80	.91	.89	.90	.92
Median	.79	.79	.79	.80	.82	.82	.83	.82

Table 4
Inter-rater Reliability Coefficients for Unweighted and Three Differential Item Procedures, by Prompt and Grade

	Prompt A				Prompt B			
Grade	Unweighted	Regression	Graded Response	Teachers' Ratings	Unweighted	Regression	Graded Response	Teachers' Ratings
2	.56	.57	.62	.58	.57	.61	.65	.59
3	.30	.35	.47	.31	.77	.78	.81	.78
4	.53	.59	.69	.52	.75	.80	.82	.77
5	.49	.54	.67	.45	.57	.63	.69	.58
6	.64	.67	.71	.65	.53	.60	.62	.55
7	.58	.61	.69	.56	.74	.79	.82	.75
8	.49	.54	.65	.47	.69	.76	.77	.69
9	.67	.68	.74	.65	.53	.63	.65	.53
10	.63	.65	.70	.61	.63	.72	.75	.62
11	.81	.83	.85	.81	.67	.75	.76	.67
12	.67	.68	.70	.68	.76	.79	.80	.78
Median	.58	.61	.69	.58	.67	.75	.76	.67

Table 5
Inter-rater Reliability Coefficients (Ebel Method) for Unweighted and Three Differential Item Weighting Procedures, by Prompt and Grade

	Prompt A				Prompt B			
Grade	Unweighted	Regression	Graded Response	Teachers' Ratings	Unweighted	Regression	Graded Response	Teachers' Ratings
2	.56	.57	.62	.58	.56	.61	.65	.59
3	.29	.34	.47	.30	.77	.78	.81	.78
4	.53	.59	.69	.52	.75	.80	.82	.77
5	.49	.54	.67	.45	.57	.63	.68	.58
6	.64	.66	.70	.65	.53	.59	.61	.54
7	.57	.60	.69	.56	.74	.79	.81	.74
8	.49	.54	.65	.47	.68	.75	.76	.67
9	.66	.67	.74	.64	.53	.63	.65	.53
10	.61	.64	.69	.59	.63	.71	.74	.62
11	.79	.82	.84	.79	.67	.75	.76	.67
12	.67	.68	.70	.68	.74	.78	.79	.75
Median	.57	.60	.69	.58	.67	.75	.76	.67